УДК 519.237.8:510.22

К.М. Залесская
ОАО «Агат-Систем», г. Минск, Беларусь
k.zaleskaya@gmail.com

Анализ устойчивости методов нечеткой кластеризации к выбору их параметров

В статье проводится анализ оптимизационных методов нечеткой кластеризации FCM, NC, PCM, FRC. Рассматриваются вопросы определения значений параметров методов нечеткой кластеризации: начальных значений центроидов и параметров доверительных границ основных кластеров. Исследуется проблема устойчивости решений задачи нечеткой кластеризации к определению значений указанных параметров.

Введение

Одной из важнейших проблем практического применения методов нечеткой кластеризации является неустойчивость их решений к наличию аномальных наблюдений в исследуемой совокупности наблюдений. Под аномальными наблюдениями здесь понимаются наблюдения, принадлежащие классам, число представителей которых в исследуемой совокупности существенно мало в сравнении с числом представителей основных классов.

Для решения задачи нечеткой кластеризации в условиях наличия в исследуемой совокупности аномальных наблюдений разработаны специальные оптимизационные методы, которые принято называть робастными [1]. Среди наиболее известных робастных методов нечеткой кластеризации необходимо выделить методы: Noise Clustering (NC) [2], Possibilistic C-Means (PCM) [3], Fuzzy Robust Clustering (FRC) [4].

Несмотря на широкое применение указанных методов, необходимо отметить недостаточное внимание, которое уделено вопросу выбора значений их параметров. Существование проблемы выбора значений параметров кластеризации для произвольной совокупности наблюдений признается исследователями и нашло отражение в немногочисленных рекомендациях [1-4].

В данной работе исследуется проблема устойчивости решений задачи кластеризации оптимизационных методов NC, PCM, FRC по отношению к определению значений параметров кластеризации: начальных значений центроидов и значений параметров доверительных границ основных классов, – а также анализируются рекомендации по определению значений указанных параметров на основе предварительного анализа исследуемой совокупности.

Проблема выбора значений параметров методов нечеткой кластеризации

Основой для разработки большинства оптимизационных методов нечеткой кластеризации послужил целевой функционал Дж. Данна и Дж. Беждека [5], лежащий в основе классического метода нечеткой кластеризации Fuzzy C-Means (FCM):

$$F_{FCM}(P) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^{\gamma} d_{ij}^{2}, \qquad (1)$$

где P – нечеткое c-разбиение исследуемой совокупности наблюдений X на заданное число c нечетких множеств A^i, $i = 1,\ldots,c$; $d_{ij} = d(x_j, \tau_i)$ – функция расстояния от наблюдения $x_j \in X$ до центроида τ_i нечеткого кластера $A^i \in \{A^1,\ldots,A^c\}$; $\gamma : 1 < \gamma < \infty$ – показатель нечеткости кластеризации; $\mu_{ij} = \mu_i(x_j) : \mu_{ij} \geq 0$ – функция принадлежности наблюдения x_j нечеткому кластеру A^i, удовлетворяющая условию:

$$\sum_{i=1}^{c} \mu_{ij} = 1 \cdot \qquad (2)$$

Попытки решения проблемы устойчивости метода FCM к аномальным наблюдениям привели к появлению методов NC, PCM, FRC. Указанные методы кластеризации допускают, что исследуемой совокупности могут принадлежать наблюдения неосновных классов, что приводит к ослаблению условия (2):

$$\sum_{i=1}^{c} \mu_{ij} \leq 1, \qquad (3)$$

где c – число основных классов, задаваемое до процедуры кластеризации.

Целевые функционалы методов NC, PCM и FRC могут быть представлены соответствующими выражениями:

$$F_{NC}(P) = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^{\gamma} d_{ij}^{2} + \sum_{j=1}^{n} \delta^{2} (1 - \sum_{i=1}^{c} \mu_{ij})^{\gamma}, \qquad (4)$$

$$F_{PCM}(P) = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij}^{\gamma} d_{ij}^{2} + \sum_{i=1}^{c} \delta_i^{2} \sum_{j=1}^{n} (1 - \mu_{ij})^{\gamma}, \qquad (5)$$

$$F_{FRC}(P) = \sum_{i=1}^{c}\sum_{j=1}^{n} \mu_{ij} d_{ij}^{2p} + \sum_{i=1}^{c}\sum_{j=1}^{n} \left(1 + \mu_{ij} * \log(\mu_{ij}) - \mu_{ij}\right) \delta_i^{2p}, \qquad (6)$$

где γ, p – показатели нечеткости кластеризации; δ, $\{\delta_i\}_{i=1}^{c}$ – параметры, определяющие доверительные границы основных кластеров.

Оптимизационную процедуру поиска экстремума вышеуказанных функционалов предваряет задание числа основных классов и выбор начальных значений центроидов, показателей нечеткости кластеризации и параметров, определяющих доверительные границы основных кластеров. Выбор значений параметров для кластеризации произвольной совокупности наблюдений является открытым вопросом нечеткой кластеризации.

Задача выбора начальных значений центроидов

Одним из важных вопросов, связанных с применением оптимизационных методов, разработанных для задач кластеризации в условиях возможного влияния аномальных наблюдений, является вопрос выбора начальных значений центроидов.

В работах [1-4], описывающих наиболее известные робастные методы нечеткой кластеризации, сообщается о важности правильного выбора начальных значений центроидов ввиду их влияния на конечные результаты кластеризации. Но сама проблема устойчивости решений задачи кластеризации к выбору начальных значений центроидов как таковая не рассматривается.

В вышеупомянутых работах даются краткие рекомендации по определению начальных значений центроидов в случае отсутствия предварительной информации о распределении наблюдений основных классов исследуемой совокупности.

В работе [1] при использовании робастных методов кластеризации предлагается выполнять итерационную процедуру оптимизационного поиска оптимального разбиения поочередно для всех возможных начальных значений центроидов, после чего осуществлять оценку полученных разбиений и выбирать на ее основании оптимальный результат кластеризации. При этом в качестве возможных начальных значений центроидов рекомендуется рассматривать все наблюдения исследуемой совокупности, хотя очевидно, что такая рекомендация неприемлема в вычислительном плане.

В работе [4] для метода FRC предлагается рассматривать не все возможные наблюдения в качестве начальных значений центроидов, а лишь наиболее удаленные друг от друга.

В работе [3] рекомендуется для определения начальных значений параметров метода PCM делать предварительную оценку исследуемой совокупности каким-либо другим методом кластеризации.

Для исследования устойчивости методов кластеризации FCM, NC, PCM, FRC к выбору начальных значений центроидов представим целевые функционалы методов как функции от центроидов $F = F(\tau)$ (где $\tau = (\tau_1, \tau_2, ...\tau_c)$ – вектор центроидов основных кластеров) при заданной исследуемой совокупности наблюдений и определенных параметрах кластеризации δ, $\{\delta_i\}_{i=1}^{c}$, γ и p с учетом оптимальности выбора параметров μ_{ij} в соответствии с выражениями из [2-5]:

$$\mu_{ij}^{FCM}(\tau) = \left(\sum_{k=1}^{c} \left[\frac{d^2(\tau_i, x_j)}{d^2(\tau_k, x_j)} \right]^{\frac{1}{\gamma-1}} \right)^{-1}; \quad (7)$$

$$\mu_{ij}^{NC}(\tau) = \left(\left(\frac{d^2(\tau_i, x_j)}{\delta^2} \right)^{\frac{1}{\gamma-1}} + \sum_{k=1}^{c} \left[\frac{d^2(\tau_i, x_j)}{d^2(\tau_k, x_j)} \right]^{\frac{1}{\gamma-1}} \right)^{-1}; \quad (8)$$

$$\mu_{ij}^{PCM}(\tau) = \left(1 + \left(\frac{d^2(\tau_i, x_j)}{\delta_i^2} \right)^{\frac{1}{\gamma-1}} \right)^{-1}; \quad (9)$$

$$\mu_{ij}^{FRC}(\tau) = \exp\left(-\left(\frac{d^2(\tau_i, x_j)}{\delta_i^2} \right)^{p} \right). \quad (10)$$

Таким образом, целевые функционалы примут вид:

$$F_{FCM}(\tau) = \sum_{j=1}^{n} \sum_{i=1}^{c} \mu_{ji}^{\gamma}(\tau) d^2(\tau_i, x_j); \quad (11)$$

$$F_{NC}(\tau) = \sum_{j=1}^{n} \sum_{i=1}^{c} \mu_{ji}^{\gamma}(\tau) d^2(\tau_i, x_j) + \sum_{j=1}^{n} \delta^2 (1 - \sum_{i=1}^{c} \mu_{ji}(\tau))^{\gamma}; \quad (12)$$

$$F_{PCM}(\tau) = \sum_{j=1}^{n} \sum_{i=1}^{c} \mu_{ji}^{\gamma}(\tau_i) d^2(\tau_i, x_j) + \sum_{i=1}^{c} \delta_i^2 \sum_{j=1}^{n} (1 - \mu_{ji}(\tau_i))^{\gamma}; \quad (13)$$

$$F_{FRC}(\tau) = \sum_{j=1}^{n} \sum_{i=1}^{c} \mu_{ji}(\tau_i) d^{2p}(\tau_i, x_j) + \sum_{i=1}^{c} \sum_{j=1}^{n} \left(1 + \mu_{ji}(\tau_i) \log(\mu_{ji}(\tau_i)) - \mu_{ji}(\tau_i)\right) \delta_i^{2p}. \quad (14)$$

Исследуем проблему выбора начальных значений центроидов на элементарном частном примере. Пусть необходимо решить задачу кластеризации совокупности наблюдений X, заданных в одномерном пространстве в соответствии с табл. 1.

Таблица 1 – Пример исследуемой совокупности наблюдений

x_1	x_2	x_3	x_4	x_5	x_6
– 0,1	– 0,05	0	0,05	0,1	2

Пусть определено число основных классов $c = 1$. Очевидно, наблюдение x_6 следует считать аномальным по отношению к основному классу, представленному наблюдениями $\{x_1, x_2, x_3, x_4, x_5\}$.

Пусть определены значения параметров кластеризации согласно табл. 2.

Таблица 2 – Пример исследуемой совокупности наблюдений

Наименование параметра	FCM	NC	PCM	FRC
Число основных классов	$c=1$	$c=1$	$c=1$	$c=1$
Доверительная граница основных кластеров	–	$\delta = 0,3$	$\delta_1 = 0,3$	$\delta_1 = 0,3$
Показатель нечеткости кластеризации	$\gamma=2$	$\gamma=2$	$\gamma=2$	$p=1$

Рассмотрим однопараметрические функции $F = F(\tau)$ (11) – (14), соответствующие каждому из методов FCM, NC, PCM и FRC, построенные для заданной совокупности наблюдений.

Функция $F_{FCM}(\tau)$ имеет один локальный экстремум в точке, смещенной относительно действительного центроида основного кластера в сторону аномального наблюдения (рис. 1 а).

В данном случае можно утверждать, что поиск оптимального разбиения методом FCM приведет к решению, соответствующему смещению центроида основного кластера относительно действительного расположения. Учитывая единственность экстремума функции $F_{FCM}(\tau)$, можно говорить об отсутствии влияния выбора начального значения центроида на результат кластеризации методом FCM. Стоит заметить, что это утверждение нельзя обобщить на общий случай для произвольного числа кластеров.

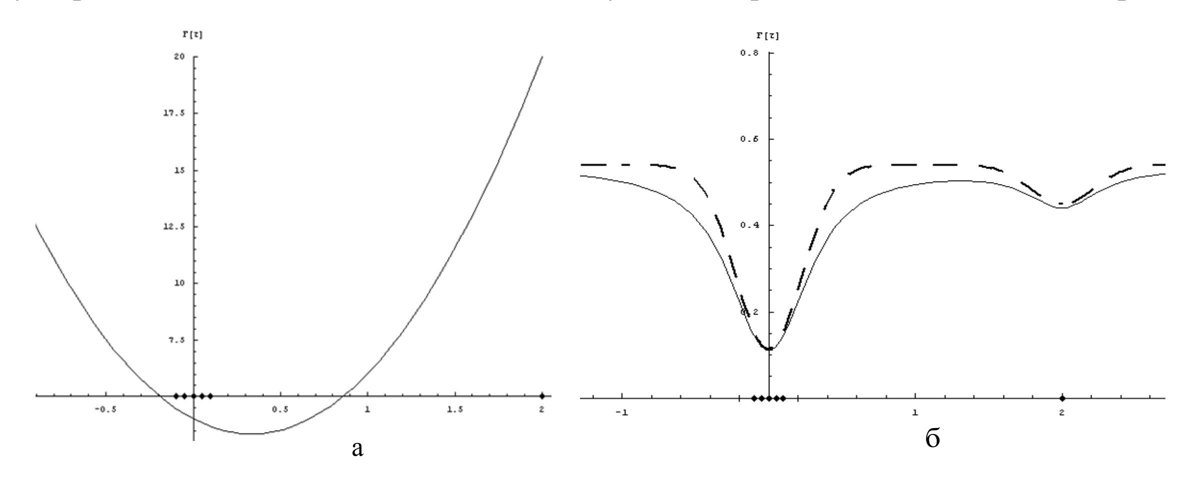


Рисунок 1 – Графики функций $F_{FCM}(\tau)$ (а);
$F_{FRC}(\tau)$ – «– – –», $F_{NC}(\tau)$ и $F_{PCM}(\tau)$ –«–» (б)

Функции $F_{NC}(\tau)$, $F_{PCM}(\tau)$, $F_{FRC}(\tau)$ имеют два локальных экстремума: в точке, близкой к действительному центроиду основного кластера, и в точке, близкой к аномальному наблюдению (рис. 1 б).

Таким образом, при использовании методов NC, PCM и FRC в качестве конечного результата кластеризации заданной совокупности может быть получено одно из двух решений: верное, соответствующее совпадению центроида с действительным центроидом основного кластера, или ложное с центроидом в точке, близкой к аномальному наблюдению. Ложное решение возможно, если начальное значение центроида принадлежит локальной области аномального наблюдения.

Данный пример показывает различие влияния выбора начального значения центроида на решения робастных методов NC, PCM и FRC в сравнении с методом FCM.

При условии корректного выбора начальных значений центроидов отклонение решений робастных методов NC, PCM и FRC от действительных центроидов основных классов в условиях аномальных наблюдений существенно мало в сравнении с решениями метода FCM. Это может быть объяснено характерным свойством целевых функционалов методов NC, PCM и FRC, которое выражается в незначительном изменении доли, вносимой аномальными наблюдениями в значения функционалов, в случае изменения положения центроидов в локальной области основных кластеров [6]. Иными словами, поиск оптимального значения центроида робастными методами осуществляется среди наблюдений локальной области заданного начального значения центроида, в отличие от метода FCM, для которого на решение задачи кластеризации одинаково влияют все наблюдения исследуемой совокупности. Это также объясняет то, что ошибочный выбор начального значения одного из центроидов в области аномального наблюдения неминуемо приводит робастные методы к ложному решению.

Рассмотренный пример показывает влияние ошибки выбора начального значения центроида на результат поиска оптимального положения отдельно взятого центроида. Учитывая особенность некоторых методов решать задачу нечеткой кластеризации при условии взаимозависимости центроидов, целесообразно исследовать влияние ошибки инициализации одного центроида на общее решение задачи кластеризации для случая нескольких кластеров.

Для этого рассмотрим следующую элементарную задачу кластеризации наблюдений в одномерном признаковом пространстве для случая двух основных кластеров. Пусть исследуемая совокупность состоит из двух наблюдений $x_1 = 0$ и $x_2 = 1$. Заданную совокупность можно считать прототипом двух компактных хорошо разделимых кластеров одинаковой мощности. Пусть определены значения параметров кластеризации согласно табл. 2.

На рис. 2 представлены функции $F_{FCM}(\tau)$, $F_{NC}(\tau)$, $F_{PCM}(\tau)$, $F_{FRC}(\tau)$ (11) – (14), характеризующие значения соответствующих целевых функционалов при различном расположении центроидов относительно фиксированных значений наблюдений.

Заметим, что все указанные функции имеют глобальные экстремумы, соответствующие действительным значениям центроидов.

Как известно, функционалы $F_{PCM}(P)$ и $F_{FRC}(P)$ представимы как суммы независимых оценок оптимальности расположения каждого отдельного центроида основного кластера [1]. Поэтому методы FRC и PCM допускают совпадение оптимальных значений центроидов $(\tau_1, \tau_2) = (x_1, x_1)$ и $(\tau_1, \tau_2) = (x_2, x_2)$.

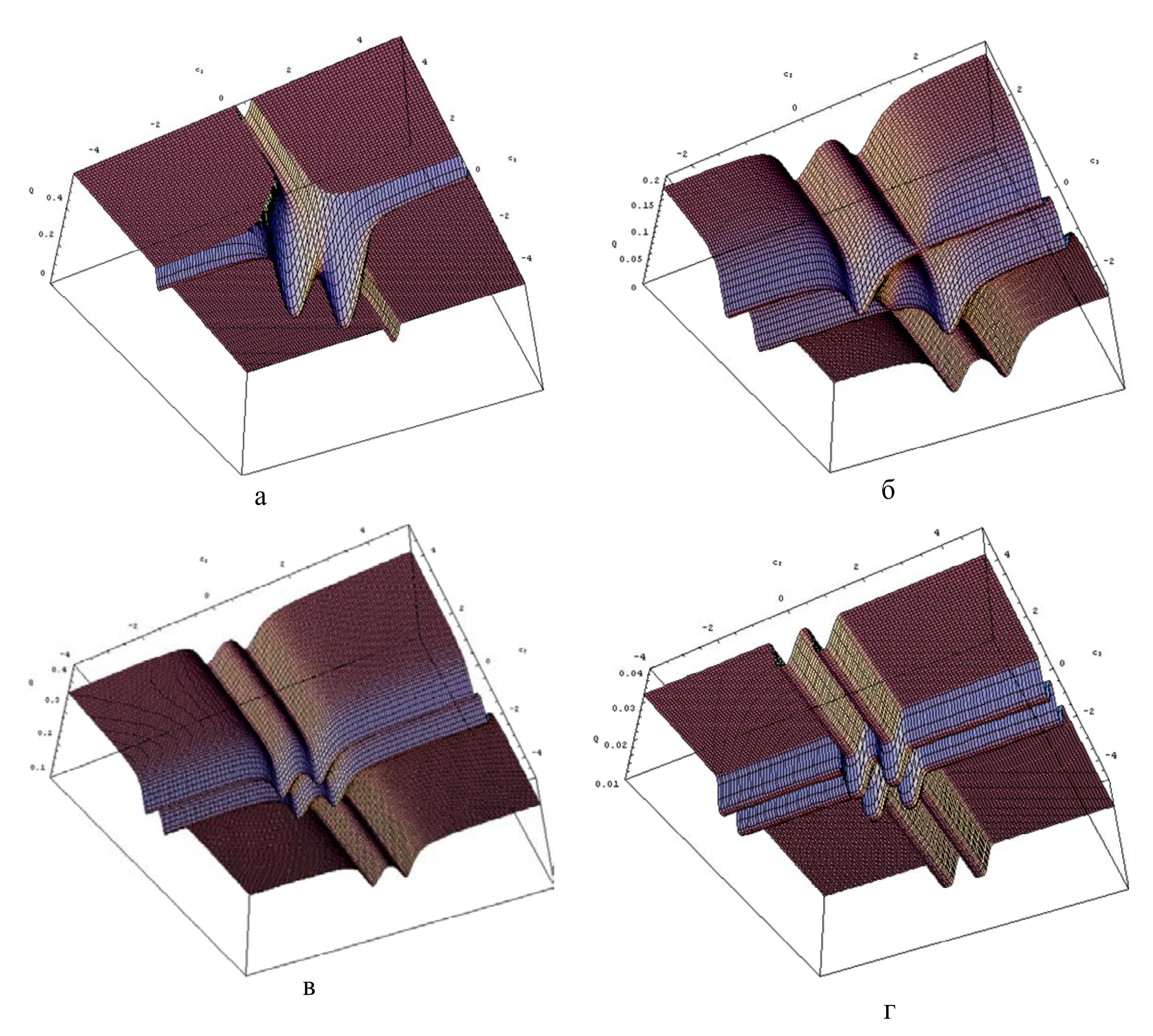


Рисунок 2 – Графики функций $F_{FCM}(\tau)$ (а), $F_{NC}(\tau)$ (б), $F_{PCM}(\tau)$ (в), $F_{FRC}(\tau)$ (г)

Пусть для центроида τ_1 найдено ложное устойчивое положение вне локальной области основных кластеров ($\tau_1 << x_1$ или $\tau_1 >> x_2$). Тогда, как следует из графиков, для методов NC, PCM и FRC оптимальным положением центроида τ_2 являются действительные центроиды основных кластеров ($\tau_2 = x_1$ и $\tau_2 = x_2$). Для методов PCM и FRC это обусловлено независимостью оценок оптимальности расположения центроидов. Заметим, что метод NC, подобно методам PCM и FRC, в случае ошибки инициализации одного из центроидов позволяет верно определять оптимальное положение остальных. Представленное свойство закономерно, поскольку влияние совокупности наблюдений на оценку оптимального положения центроида ограничено локальной областью. При тех же условиях для метода FCM оптимальным положением центроида τ_2 является точка, равноудаленная от действительных центроидов основных кластеров: $\tau_2 = (x_1 + x_2)/2$. Таким образом, ошибка выбора начального значения одного центроида приводит к ошибке решения задачи кластеризации в целом.

Вышеприведенные примеры позволяют сформировать представление о последствиях ошибочного выбора начальных значений центроидов.

Пользуясь выводами проведенного исследования, можно утверждать, что некоторые рекомендации [1-4] по выбору начальных значений центроидов для робастных методов являются в общем случае неправомерными. При выборе начальных значений

центроидов среди наиболее удаленных друг от друга наблюдений, либо среди произвольных наблюдений исследуемой совокупности, велика вероятность выбора аномального наблюдения в качестве центроида, что может привести к нахождению ложного разбиения. Многократное решение задачи кластеризации при переборе значений параметров инициализации трудоемко и не всегда доступно исследователю.

Рекомендацию использования процедуры предварительной оценки исследуемой совокупности с помощью альтернативного метода кластеризации, не обязательно нечеткого, для определения начальных значений центроидов, приведенную в [3], следует считать применимой не только для метода PCM, но и для методов NC и FRC. При использовании метода FCM для предварительной оценки необходимо помнить о чувствительности решений данного метода к аномальным наблюдениям.

Задача выбора значений параметров доверительных границ основных кластеров

Целевой функционал $F(P)$ методов NC, PCM, FRC может быть охарактеризован показательной функцией $Q = Q(x)$, значение которой в произвольной точке x является некоторой долей, вносимой наблюдением x в общую оценку разбиения P при условии принадлежности x исследуемой совокупности [6]. Функции $Q(x)$ методов NC, PCM, FRC ограничены во всем признаковом пространстве и имеют асимптотическое поведение вне локальных областей заданных центроидов основных кластеров. Данное свойство позволяет ограничить влияние на оценку разбиения аномальных наблюдений, находящихся вне определенных локальных областей центроидов основных кластеров. Размер каждой из таких локальных областей центроидов определяется соответствующим параметром доверительной границы основного кластера, что объясняет чувствительность решения задачи кластеризации к выбору значений этих параметров.

Обобщим рекомендации по выбору значений параметров доверительных границ основных кластеров δ, $\{\delta_i\}_{i=1}^{c}$, приведенные в работах [1-4], [7]. В соответствии с указанными источниками значения параметров δ, $\{\delta_i\}_{i=1}^{c}$ должны быть установлены на этапе, предваряющем процедуру кластеризации методами NC, PCM и FRC. В процессе кластеризации значения выбранных изначально параметров δ, $\{\delta_i\}_{i=1}^{c}$ не изменяются или корректируются.

Согласно работе [2] параметр δ для метода NC предлагается выбирать в соответствии с выражением:

$$\delta = \frac{\lambda}{nc}\sum_{j}^{n}\sum_{i}^{c} d(x_j, \tau_i), \qquad (15)$$

где λ – масштабирующий параметр; c – число кластеров; τ_i – центроид нечеткого кластера $A^i \in \{A^1, \ldots, A^c\}$; n – количество наблюдений исследуемой совокупности. В данной рекомендации для определения значений параметров λ и $\{\tau_i\}_{i=1}^{c}$ требуется оценка исследуемой совокупности, предваряющая инициализацию параметра δ.

Авторы работы [7] предлагают избежать предварительного определения значений центроидов $\{\tau_i\}_{i=1}^{c}$, используя для выбора значения параметра δ выражение:

$$\delta = \alpha r, \qquad (16)$$

где α – масштабирующий параметр, r – радиус c равновеликих гиперобъемов минимального размера, выделенных в признаковом пространстве и способных вместить все наблюдения исследуемой совокупности. Для использования данной рекомендации исследователю необходимо самостоятельно определить алгоритм выделения гиперобъема.

В работах [1], [3], [4] для выбора значений параметров $\{\delta_i\}_{i=1}^{c}$ методов PCM и FRC предлагается использовать результаты предварительной кластеризации. В частности, начальные значения $\{\delta_i\}_{i=1}^{c}$ рекомендуется определять как взвешенные внутриклассовые расстояния соответствующих кластеров $\{A^i\}_{i=1}^{c}$:

$$\delta_i = \frac{\lambda}{n_i} \sum_{x_j \in (P_i)_\alpha} \omega_{ij} d^2(x_j, \tau_i), \tag{17}$$

где λ – масштабирующий параметр; n_i – нормирующий коэффициент, оценивающий количество наблюдений нечеткого кластера A^i, в качестве которого могут выступать $\left|(P_i)_\alpha\right|$ – мощность подмножества α -уровня нечеткого c-разбиения P исследуемой совокупности наблюдений X, либо $\sum_{j=1}^{n} \mu_{ij}^{\gamma}$ – сумма степеней принадлежности наблюдений кластеру A^i; $\omega_{ij}, \omega_j \in [0, 1]$ – весовой коэффициент, характеризующий оценку принадлежности наблюдения x_j нечеткому кластеру A^i, в качестве которого могут использоваться значения степеней принадлежности μ_{ij}.

В работе [4] предлагается выбирать начальные значения параметров $\{\delta_i\}$ функционалов $F_{PCM}(P)$ и $F_{FRC}(P)$ в соответствии с выражением:

$$\delta_i = \min_k \left\{\left\|\tau_k - \tau_i\right\|\right\}, \quad k \neq i, \tag{18}$$

причем в качестве начальных значений центроидов $\{\tau_i\}_{i=1}^{c}$ допускается выбирать наблюдения, находящиеся на наибольшем удалении друг от друга.

Для оценки данных рекомендаций проведем исследование устойчивости робастных методов кластеризации к выбору значений параметров доверительных границ основных кластеров.

Представим целевые функционалы $F_{PCM}(P)$ и $F_{FRC}(P)$ как суммы компонент $R(\tau_i, \delta_i)$, $i = 1, \ldots, c$, каждый из которых является оценкой оптимальности расположения центроида τ_i с соответствующим ему параметром δ_i по отношению к исследуемой совокупности наблюдений:

$$F(P) = \sum_{i=1}^{c} R(\tau_i, \delta_i), \tag{19}$$

$$R_{PCM}(\tau, \delta) = \sum_{j=1}^{n} \mu_j^{\gamma}(\tau) d^2(\tau, x_j) + \delta^2 \sum_{j=1}^{n} (1 - \mu_j(\tau))^{\gamma}, \tag{20}$$

$$R_{FRC}(\tau, \delta) = \sum_{j=1}^{n} \mu_j(\tau) d^{2p}(\tau, x_j) + \delta^{2p} \sum_{j=1}^{n} \left(1 + \mu_j(\tau)\log(\mu_j(\tau)) - \mu_j(\tau)\right). \tag{21}$$

Исследуем проблему выбора параметров доверительной границы основных кластеров на элементарном частном примере. Пусть необходимо решить задачу кластери-

зации совокупности наблюдений X, заданных в одномерном пространстве в соответствии с табл. 3.

Таблица 3 – Пример исследуемой совокупности наблюдений

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
–0,125	–0,075	–0,025	0,025	0,075	0,125	0,85	0,9	0,95	1,05	1,1	1,15

Исследуемая совокупность X (табл. 3) представлена двумя хорошо разделимыми классами одинаковой мощности с центрами в точках 0 и 1.

На рис. 3 представлены графики функций $R_{PCM}(\tau,\delta)$ и $R_{FRC}(\tau,\delta)$ при разных значениях параметра δ .

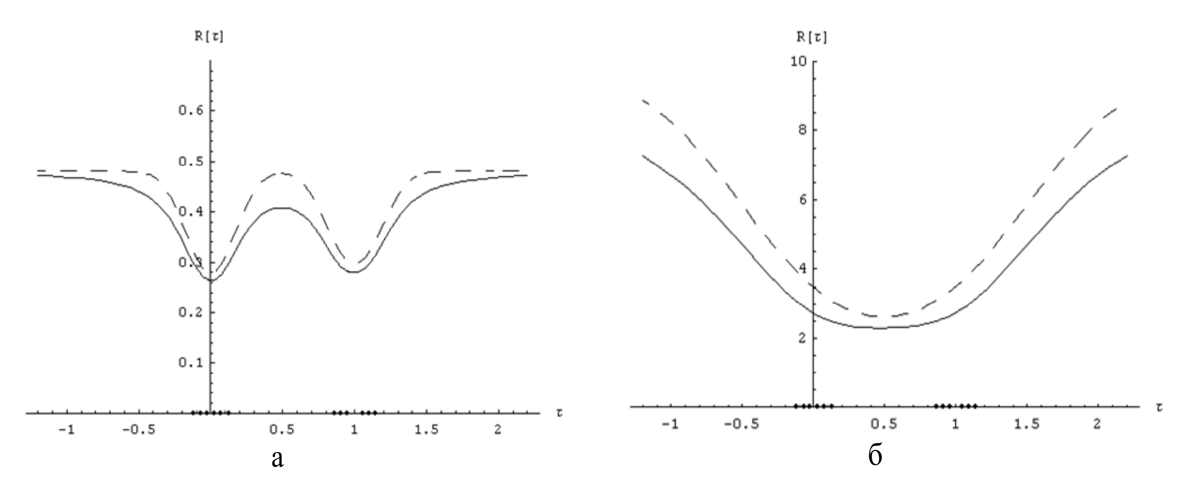


Рисунок 3 – Графики функций R_{PCM} «–», R_{FRC} «– –»
при значении параметра $\delta = 0{,}2$ (а) и $\delta = 0{,}9$ (б)

Графики показывают, что в зависимости от выбора значения параметра δ , функции $R_{PCM}(\tau,\delta)$ и $R_{FRC}(\tau,\delta)$ могут иметь различное число локальных экстремумов.

В случае выбора значения параметра δ , близкого к значению расстояния от действительного центроида основного кластера до наиболее удаленного наблюдения этого кластера, функции $R_{PCM}(\tau,\delta)$ и $R_{FRC}(\tau,\delta)$ имеют два локальных экстремума в точках, близких к центрам классов.

При выборе значения параметра δ , превышающего половину расстояния между центрами классов, функции $R_{PCM}(\tau,\delta)$ и $R_{FRC}(\tau,\delta)$ вырождаются в функции, имеющие один локальный экстремум. Это может быть объяснено пересечением доверительных границ основных кластеров и их объединением в один кластер. При таком выборе значения δ оптимальным расположением центроида является точка, находящаяся между действительными центрами кластеров.

При выборе значения параметра δ значительно меньше значения расстояния от действительного центроида основного кластера до наиболее удаленного наблюдения этого кластера вероятно появление нескольких экстремумов функций $R_{PCM}(\tau,\delta)$ и $R_{FRC}(\tau,\delta)$, соответствующих одному и тому же основному кластеру (рис. 4). Такое разделение кластера возможно в случае, если наблюдения распределены по закону, отличному от нормального.

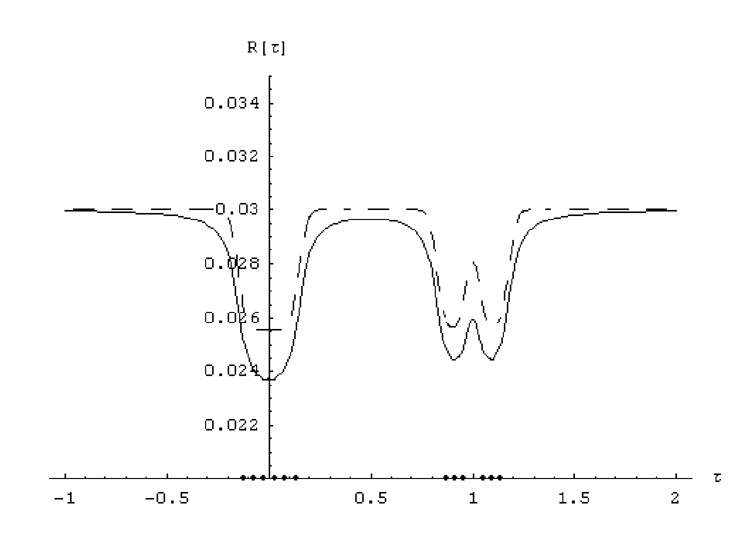


Рисунок 4 – Графики функций R_{PCM} «–», R_{FRC} «– –» при значении параметра $\delta = 0{,}05$

Вышеприведенные примеры позволяют сформировать представление о последствиях ошибочного выбора параметров доверительных границ основных кластеров.

Отметим еще одну важную особенность робастных методов PCM и FRC. Упомянутые методы кластеризации допускают выбор различных параметров доверительных границ для различных основных кластеров. Пусть для двух кластеров A_{i_1} и A_{i_2} определены соответственно значения параметров доверительных границ δ_{i_1} и δ_{i_2}, причем $\delta_{i_1} > \delta_{i_2}$. Тогда для любой невырожденной выборки для любых значений τ справедливо неравенство:

$$R(\tau, \delta_{i_1}) > R(\tau, \delta_{i_2}).$$

Это свойство функций $R_{PCM}(\tau, \delta)$ и $R_{FRC}(\tau, \delta)$ является следствием выражений (20) и (21). Данное свойство исключает возможность сравнения оценок разбиения $F_{PCM}(P)$ и $F_{FRC}(P)$, полученных при выборе различных значений параметров $\{\delta_i\}$.

Результаты проведенного исследования позволяют выявить характер последствий ошибочного выбора значений параметров доверительных границ основных кластеров. Заметим неправомерность некоторых рекомендаций.

При выборе значения каждого из параметров δ_i, $i = 1, \ldots, c$ как расстояния между центроидами $\min_i \{\|\tau_k - \tau_i\|\}$, $k \neq i$, согласно [4], велика вероятность объединения кластеров A^k и A^i, что не соответствует действительному разбиению.

Изменение значений параметров $\{\delta_i\}$ согласно [3] в рамках многократного выполнения процедуры оптимизационного поиска решения не позволяет оценивать результаты кластеризации, непосредственно сравнивая значения функционалов. Для сравнения должны быть использованы иные критерии оценки разбиения.

Для правильного выбора значений параметров δ, $\{\delta_i\}$ методов NC, PCM и FRC обязательным условием должно быть наличие некоторой предварительной оценки распределения каждого основного класса исследуемой совокупности.

Представленные в [2], [3] рекомендации выбора значений параметров δ, $\{\delta_i\}$, основанные на предварительной оценке совокупности наблюдений с помощью альтернативного метода кластеризации, следует считать применимыми не только для методов NC и PCM, но и для метода FRC.

Выводы

Как известно, наличие аномальных наблюдений в выборке приводит к существенным ошибкам решений задачи нечеткой кластеризации с использованием классического метода FCM. Решение проблемы может быть найдено в применении робастных методов кластеризации, среди которых наиболее известными являются методы NC, PCM и FRC. Следует заметить, что ослабление влияния аномальных наблюдений на решения задачи кластеризации методами NC, PCM и FRC достигается с помощью выделения некоторых доверительных областей основных кластеров. Наблюдения, находящиеся вне доверительных областей основных кластеров, вносят приблизительно одинаковую долю в общую оценку разбиения и не оказывают значительного влияния на результаты кластеризации. Применение указанных робастных методов нечеткой кластеризации осложняется проблемой определения значений их параметров, характеризующих доверительные области основных кластеров.

В данной работе проведен анализ влияния ошибки определения параметров методов кластеризации NC, PCM и FRC: начальных значений центроидов и параметров доверительных границ основных кластеров. Проведено сравнение устойчивости решений робастных методов кластеризации NC, PCM и FRC и метода FCM к выбору начальных значений центроидов. На примере методов PCM и FRC исследовано влияние ошибки определения значений параметров доверительных границ основных кластеров.

На основании частных элементарных примеров сделан вывод о неправомерности некоторых авторских рекомендаций упрощенного выбора значений указанных параметров.

В рамках исследований была установлена необходимость предварительной оценки распределения основных классов исследуемой совокупности наблюдений для определения значений параметров робастных методов NC, PCM и FRC и возможность использования с этой целью методов кластеризации.

Литература

1. Davé R.N. Robust Clustering Methods: A Unified View / R.N. Davé, R. Krishnapuram // IEEE Transactions on Fuzzy Systems. – 1997. – № 5. – P. 270-293.
2. Dave R.N. Characterization and detection of noise in clustering / R.N. Davé // Pattern Recognition. – 1991. – Vol. 11, № 12. – P. 657-664.
3. Krishnapuram R. A possibilistic approach to clustering / R. Krishnapuram, J.M. Keller // IEEE Trans. Fuzzy Systems. – 1993. – № 1 – P. 98-110.
4. Yang T.-N. Competitive algorithm for the clustering of noisy data / T.-N. Yang, S.-D. Wang // Fuzzy Sets and Systems. – 2004. – № 141. – P. 281-299.
5. Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms / Bezdek J.C. – New York. : Plenum Press, 1981. – 230 p.
6. Садовская К.М. Анализ устойчивости методов нечеткой кластеризации к аномальным наблюдениям / К.М. Садовская // Информатика. – 2009. – 4. – № 24.
7. Rehm F. A novel approach to noise clustering for outlier detection / F. Rehm, F. Klawonn, R. Kruse // Soft Comput. – 2007. – № 11. – P. 489-494.

К.М. Залеська

Аналіз стійкості методів нечіткої кластеризації щодо вибору їх параметрів

Проводиться аналіз оптимізаційних методів нечіткої кластеризації FCM, NC, PCM, FRC. Розглядаються питання щодо визначення значень параметрів методів нечіткої кластеризації: початкових значень центроїдів і параметрів довірчих меж основних кластерів. Досліджується проблема стійкості рішень задачі нечіткої кластеризації відносно визначення значень вказаних параметрів.

K.M. Zaleskaya

The Analysis of Fuzzy Clustering Methods Stability to the Choice of their Parameters

The analysis of optimization of fuzzy clustering methods FCM, NC, PCM, FRC is being made. The question of definition of parameters values of fuzzy clustering method such as initial values of centroids and the parameters of the primal cluster confidence limits is being considered. The question of the stability of fuzzy clustering problem solution with reference to the definition of the parameters specified values is being researched.

Статья поступила в редакцию 01.06.2010.